processes

MDPI

Article

Forecasting of Air Quality Using an Optimized Recurrent Neural Network

Khawaja Hassan Waseem [1], Hammad Mushtaq [1], Fazeel Abid [1], Adnan M. Abu-Mahfouz [2,3], Asadullah Shaikh [4], Mehmet Turan [5] and Jawad Rasheed [6,*]

[1] Department of Information Systems, University of Management and Technology, Lahore 54770, Pakistan
[2] Council for Scientific and Industrial Research (CSIR), Pretoria 0184, South Africa
[3] Department of Electrical and Electronic Engineering Science, University of Johannesburg, Johannesburg 2006, South Africa
[4] College of Computer Science and Information Systems, Najran University, Najran 61441, Saudi Arabia
[5] Institute of Biomedical Engineering, Boğaziçi University, Istanbul 34342, Turkey
[6] Department of Software Engineering, Nisantasi University, Istanbul 34398, Turkey
* Correspondence: jawad.rasheed@nisantasi.edu.tr

Abstract: Clean air is necessary for leading a healthy life. Many respiratory illnesses have their root in the poor quality of air across regions. Due to the tremendous impact of air quality on people's lives, it is essential to devise a mechanism through which air pollutants ($PM_{2.5}$, NO_x, CO_x, SO_x) can be forecasted. However, forecasting air quality and its pollutants is complicated as air quality depends on several factors such as weather, vehicular, and power plant emissions. This aim of this research was to find the impact of weather on $PM_{2.5}$ concentrations and to forecast the daily and hourly $PM_{2.5}$ concentration for the next 30 days and 72 h in Pakistan. This forecasting was done through state-of-the-art deep learning and machine learning models such as FbProphet, LSTM, and LSTM encoder–decoder. This research also successfully forecasted the proposed daily and hourly $PM_{2.5}$ concentration. The LSTM encoder–decoder had the best performance and successfully forecasted $PM_{2.5}$ concentration with a mean absolute percentage error (MAPE) of 28.2%, 15.07%, and 42.1% daily, and 11.75%, 9.5%, and 7.4% hourly for different cities in Pakistan. This research proves that a data-driven approach is essential for resolving air pollution in Pakistan.

Keywords: air quality; forecasting; $PM_{2.5}$; forecasting; time series models; FbProphet; neural network

Citation: Waseem, K.H.; Mushtaq, H.; Abid, F.; Abu-Mahfouz, A.M.; Shaikh, A.; Turan, M.; Rasheed, J. Forecasting of Air Quality Using an Optimized Recurrent Neural Network. *Processes* **2022**, *10*, 2117. https://doi.org/10.3390/pr10102117

Academic Editors: Marcin Banach, Olga Długosz and Jolanta Pulit-Prociak

Received: 28 August 2022
Accepted: 12 October 2022
Published: 18 October 2022

1. Introduction

Clean air is paramount for healthy human life, thus making air quality maintenance an integral part of public health policy. However, in recent years due to increasing urbanization, industrialization, and deforestation, the issue of air pollution is becoming more and more potent. Air pollution is caused primarily due to the introduction of harmful chemical, biological, and particulate matter into our atmosphere. Among these dangerous materials, the most common and abundant is particulate matter 2.5 ($PM_{2.5}$). This fine particulate matter is composed of a mixture of solid and liquid particles in the air. Their abundance above a certain threshold leads to smog and a hazy environment. When inhaled into the human body, these result in various cardiac and pulmonary problems. There exists a correlation between air pollution and meteorological conditions. Factors such as wind, rain, temperature, pressure, ultraviolet radiation, and humidity can impact air pollution in a region. Therefore, a thorough understanding of the weather is pertinent when analyzing a region's air quality properly.

Pakistan especially has suffered the full brunt of this crisis. Many megacities have been suffering from smog and haze, resulting in various health problems for the residents.

According to the World Health Organization, the air quality inside Pakistan is generally considered unsafe. The most recent data published by the WHO indicates that $PM_{2.5}$ concentration across the region is, on average, 58 μg/m³, which is higher than the prescribed safety factor of 10 μg/m³ [1]. $PM_{2.5}$ concentration is exceptionally high in urban centers such as Lahore, Karachi, and Islamabad. In 2019, with an extremely high $PM_{2.5}$ concentration of 68 μg/m³, Pakistan was declared the world's 2nd most polluted country by the world air quality report published by IQAir [2]. Figures 1–3 depict the $PM_{2.5}$ concentration from 2019 to 2021.

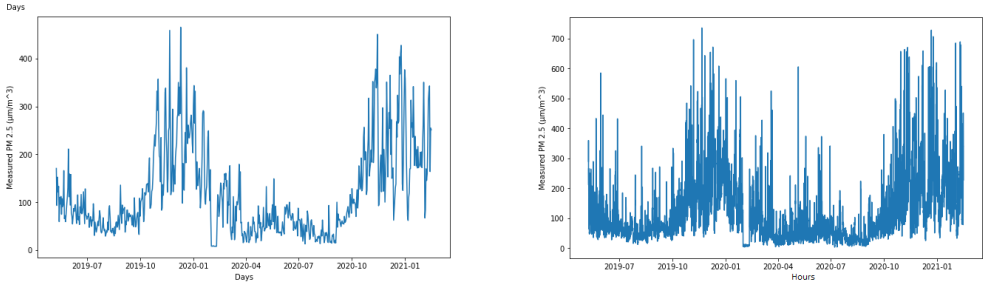

Figure 1. Lahore daily and hourly $PM_{2.5}$ concentrations between 2019–2021.

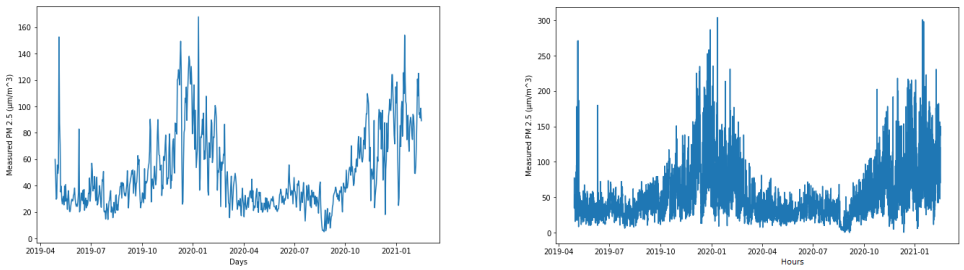

Figure 2. Islamabad daily and hourly $PM_{2.5}$ concentrations between 2019–2021.

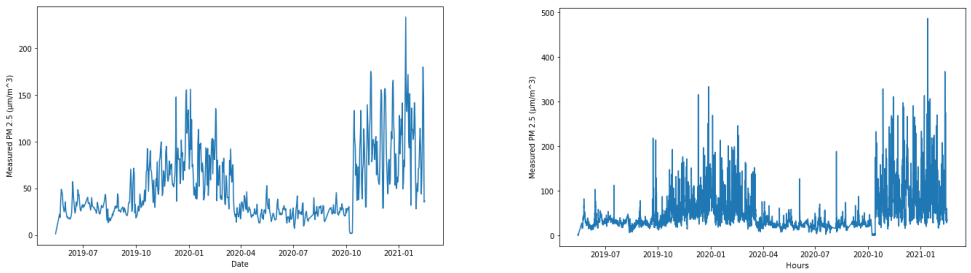

Figure 3. Karachi daily and hourly $PM_{2.5}$ concentrations between 2019–2021.

An efficient and streamlined monitoring system needs to be developed through which the government can record and forecast air quality levels. This intervention would allow the local populace to take precautionary measures in the event of deterioration in air quality levels. At the same time, it would help the government bodies in terms of evidence-based policymaking for air pollution abatement. However, such a robust system requires a highly efficient forecasting model, which can accurately predict air quality over time. We tried to fill this gap using machine learning algorithms to develop an air quality prediction model. The main contributions of this work are as follows.

- Data on air quality, air pollutants ($PM_{2.5}$), and meteorological conditions for multiple cities in Pakistan were combined to produce a novel dataset.
- The impact of different meteorological conditions such as temperature, humidity, precipitation, wind speed, dew point, and pressure on the daily $PM_{2.5}$ concentration in multiple Pakistani cities was found.

- Several machine and deep learning models, including multivariate FbProphet, LSTM, and LSTM encoder–decoder, were used for the daily and hourly forecasting of PM2.5 levels across numerous cities in Pakistan.

2. Related Work

Various Environmental Protection Agencies (EPA) have offered a variety of methodologies for calculating the air quality index. While most of these agencies have shifted toward the state of art machine learning techniques for forecasting the AQI, many agencies still rely on mathematical calculation. In [3], the author employed the popular machine learning method of support vector regression (SVR) to forecast the pollutant and particulate levels and predict the resulting value of the AQI in California, USA. The author employed the radial base function (RBF) and SVR to obtain the most accurate prediction. In the six AQI categories defined by the US Environmental Protection Agency, the proposed model was able to perform at a high accuracy of 94.5%. The proposed approach's limitation was its limited amount of data and parameters, especially for NO2 and $PM_{2.5}$. In [4], the author suggested an AQI and NOx forecasting method using SVR and the random forest method. Their proposed study showed that the SVR-based model performed better than the random forest model for forecasting AQI and NOx. In [5], the author proposed using multinominal regression and K nearest neighbor to predict different AQI buckets. These buckets contained the overall classification of the AQI as good, moderate, and severe.

Environmental Protection Agencies have proposed various methods for measuring the air quality index. While most of these agencies have shifted towards state-of-the-art machine learning techniques for forecasting the AQI, many agencies still rely on mathematical calculation.

In [6], the author used the previous day's temperature, humidity, dew point, wind speed, pressure, visibility, and precipitation as predictors in their ANFIS model. The author employed techniques such as collinearity tests and forward selection (FS) to minimize the cost and time of calculation. These techniques removed the redundant input variables and selected different input variable combinations. This method produced a different model for the different constituents of pollutant prediction with a better accuracy and reduced the computational time.

In [7], the author proposed the use of hybrid single decomposition (HSD) and hybrid two-phase decomposition (HTPD) for predicting the AQI a day before the next day in advance. Among all the models, the performance of HTSD was the most accurate. Their model successfully reduced the raw data instability and simplified the intrinsic complexities of daily AQI prediction.

In [8], the author found a clear relationship between visibility and the AQI. The author concluded that the AQI and image visibility were negatively correlated. As visibility increased, the AQI value decreased and vice versa. The author employed these images with high and low $PM_{2.5}$ concentrations to obtain high-frequency information. The SVR model was then updated using this data. This approach provided a rapid and cost-effective method for the prediction of the AQI. In [9], the author conducted a comparative analysis of air quality in Taiwan and London. The author analyzed the air quality in multiple stations in Taiwan and proposed an enhanced decision tree model that could predict air quality levels with an R^2 value of 0.71 and root mean squared error (RMSE) of 7.06.

To improve the accuracy of forecasting air pollutants in Shenzen, China, ref. [10] proposed a hybrid method consisting of ARIMA and the prophet method. They applied this hybrid model to 11 stations in the city and performed an error evaluation. They found that these hybrid methods improved the prediction result. However, the proposed method's processing speed was slow compared to other machine learning approaches. Similarly, in [11], the author proposed using the ARIMA model to forecast the air pollutant (NO_x, SO_2, SPM, and RSPM) levels for the next five years in Nanded city, Maharastra, India.

It is necessary to conduct a complete analysis of all the factors that influence the air pollution in a region. However, most research has been limited to the relationship between the weather and air pollution. In contrast to conventional methods, ref. [12] used XGboost and Bayesian optimization to investigate environmental, demographic, economic, and meteorological causes. This case study, which was conducted in the USA, provided excellent results.

In [13], the author proposed forecasting air quality for the next 48 h using a combination of neural network models. These models included artificial neural networks (ANN), convolutional neural networks (CNN), and long short-term memory (LSTM). They made use of this model to extract spatial–temporal relations. Their model outperformed many state-of-the-art models. The only weakness in their approach was the noise in their data. This noise was due to the use of different machines for data collection, which decreased the accuracy of the results.

In [14], the author proposed using a hybrid ensemble model, CERL, for forecasting the hourly air quality in northwest China. The advantage of using CERL was that it exploited the benefits of both feed-forward and recurrent neural networks. Through this model, they forecasted the air pollutants from 1 to 8 h ahead with a relatively high accuracy from 1 to 20% concerning the step size.

In [15], the author used a novel deep learning method for forecasting $PM_{2.5}$ concentrations in Beijing, China. The proposed architecture consisted of a hybrid deep learning model, which was a combination of a one-dimensional convolutional neural network (1D-CNN) and a bi-directional long short-term memory (Bi-LSTM). A 1D-CNN was used to extract local trends and spatial features, while a Bi-LSTM was used to learn spatial–temporal dependencies. The author conducted extensive experiments and achieved a satisfactory accuracy with this model.

Air pollutant concentration is dependent on various factors. These factors are usually either left unexplored or used in their entirety to forecast pollutants. In [16], the author suggested a new feature extraction method for air pollutant prediction, especially for $PM_{2.5}$. The author proposed a causality-based linear method to extract the most relevant features for predicting $PM_{2.5}$. Their findings proved that the proposed feature extraction had vastly improved prediction results.

In [17], the author proposed using a combination of RNN and LSTM to forecast O_3 levels for the next 8 to 72 h. The author used a decision tree to identify input variables of the highest importance and then used these features for training the model. The proposed model was able to achieve a satisfactory accuracy, and the mean absolute error was less than 2 for the 72-h sequence for forecasting. The disadvantage of this approach was that it utilized a limited number of features for training the model, which might lead to optional bias in the results. Similarly, the author analyzed the AQI and $PM_{2.5}$ concentration in the Chinese city of Fuzhou. The author applied the ARIMA model to analyze and forecast the $PM_{2.5}$ concentration between 2014–2016. The results of the study concluded that the $PM_{2.5}$ concentration had an intricate relation between seasons and that the concentration was sufficiently more significant in winter compared to summer. This study was unique as it was conducted on new data as well as it being able to analyze the seasonality of $PM_{2.5}$ over time.

This study consists of five sections. Section 2 discusses the methodology employed for forecasting the hourly and daily forecasting of $PM_{2.5}$. Section 3 presents the results and analyzes them. Section 4 discusses the research findings. Finally, Section 5 concludes the study and discusses the limitations and future work in this domain.

3. Materials and Methods

This section describes the system architecture and all the models used to solve the problems highlighted in the objectives. This section consists of 4 subsections. Section 3.1 defines the system architecture for the solution of the problem state. Section 3.2 explains

the process for the collection of data from multiple sources. Section 3.3 deals with the preprocessing of data. Section 3.4 describes the methodology of implementing proposed models to obtain the required forecast.

3.1. Model Architecture

The model architecture describes the overall steps taken to derive results from a system. Figure 4 depicts the model architecture for the proposed system. In the first step, the air quality and weather data are collected from the required sources, as illustrated in Figure 4. Then, this data undergoes the required data preprocessing and feature engineering. Then, the data is split and scaled and passed over the proposed models, i.e., multivariate FbProphet, LSTM, and LSTM encoder and decoder, to forecast the future $PM_{2.5}$ levels. Finally, the proposed models are tested and compared after undergoing hyper parameterization.

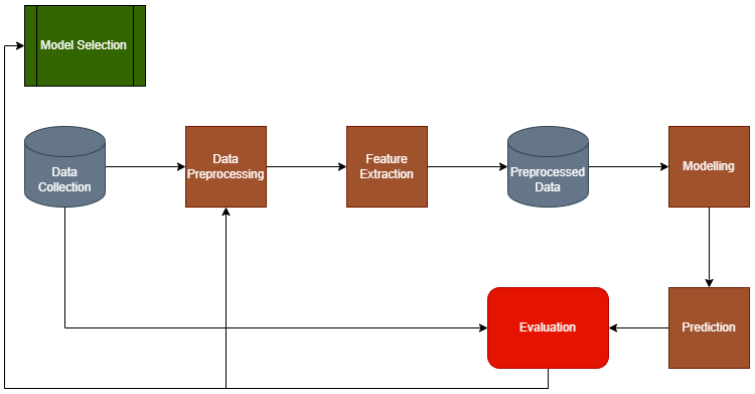

Figure 4. Model architecture for the analysis.

3.2. Data Collection

The data in this study was collected primarily from two sources: air quality data from sensors located at US embassies across Pakistan [18] and meteorological data from the World Weather website [19]. Data was recorded at hourly intervals from mid 2019 to Feb 2021. The meteorological data consisted of hourly features such as time, humidity, temperature, precipitation (in inches), UV index, wind speed, cloud cover, visibility, dew point, and pressure. Similarly, the air quality consisted of hourly features such as time, AQI, and $PM_{2.5}$ concentration. Table 1 displays the statistical distribution of the Lahore, Islamabad, and Karachi datasets.

The dataset was divided into training and testing datasets. For the daily $PM_{2.5}$ prediction, the training dataset was from mid 2019 to Jan 2019, while the testing dataset was from Jan 2021 to Feb 2021 (30 Days). Similarly, for the hourly $PM_{2.5}$ prediction, the training dataset was from mid 2019 to Feb 2019, while the testing dataset was from the last 72 h (72 H).

Table 1. Statistical distribution of the datasets.

Sr. No.	City	Features	Count	Mean	Std	Min	25%	50%	75%	Max	Unique	Top	Freq.
1	Lahore	Temp (C)	15,500.0	28.7	9.8	5.0	20.0	30.0	37.0	52.0			
		Wind speed (kmph)	15,500.0	8.6	4.2	0.0	6.0	8.0	11.0	40.0			
		Precipitation (mm)	15,500.0	0.1	0.3	0.0	0.0	0.0	0.0	15.6			
		Humidity	15,500.0	37.4	19.6	4.0	22.0	34.0	50.0	97.0			
		Visibility (km)	15,500.0	10.1	1.5	2.0	10.0	10.0	10.0	20.0			
		Pressure (in)	15,500.0	30.1	0.3	30.0	30.0	30.0	30.0	31.0			
		Cloud cover	15,500.0	20.8	25.5	0.0	1.0	8.0	34.0	100.0			

		Dew point (C)	15,500.0	10.1	8.4	−13.0	4.0	10.0	17.0	26.0		
		UV index	15,500.0	4.0	3.5		1.0	1.0	7.0	11.0		
		PM2.5 conc.	15,500.0	118.9	108.3	2.6	44.3	78.4	161.6	735.5		
		Is daytime	15,500.0								2.0	no 8405.0
		Wind direction (16 points)	15,500.0								16.0	ESE 1994.0
		Weather description	15,500.0								21.0	Partly cloudy 8836.0
2	Islamabad	Temp (C)	15,769.0	25.1	9.3	4.0	17.0	26.0	32.0	48.0		
		Wind speed (kmph)	15,769.0	8.8	3.4	1.0	6.0	8.0	10.0	29.0		
		Precipitation (mm)	15,769.0	0.1	0.5	0.0	0.0	0.0	0.0	13.8		
		Humidity	15,769.0	40.5	17.8	6.0	27.0	38.0	53.0	95.0		
		Visibility (km)	15,769.0	10.0	1.9	2.0	10.0	10.0	10.0	20.0		
		Pressure (in)	15,769.0	30.2	0.4	30.0	30.0	30.0	30.0	31.0		
		Cloud cover	15,769.0	28.8	31.2	0.0	2.0	13.0	58.0	100.0		
		Dew point (C)	15,769.0	9.0	8.5	−12.0	3.0	9.0	16.0	25.0		
		UV index	15,769.0	3.6	3.1	1.0	1.0	1.0	7.0	11.0		
		PM2.5 conc.	15,769.0	49.7	36.9	0.0	25.9	37.5	61.2	304.0		
		Is daytime	15,769.0								2.0	no 8550.0
		Wind direction (16 points)	15,769.0								16.0	NE 1871.0
		Weather description	15,769.0								20.0	Partly cloudy 7931.0
3	Karachi	Temp (C)	14,852.0	27.7	4.6	12.0	25.0	29.0	31.0	41.0		
		Wind speed (kmph)	14,852.0	19.8	8.3	1.0	13.0	19.0	25.0	52.0		
		Precipitation (mm)	14,852.0	0.0	0.2	0.0	0.0	0.0	0.0	8.7		
		Humidity	14,852.0	53.3	21.4	7.0	34.0	58.0	72.0	92.0		
		Visibility (km)	14,852.0	10.2	1.4	3.0	10.0	10.0	10.0	20.0		
		Pressure (in)	14,852.0	30.1	0.3	30.0	30.0	30.0	30.0	31.0		
		Cloud cover	14,852.0	21.2	28.5	0.0	0.0	5.0	38.0	100.0		
		Dew Point (C)	14,852.0	15.9	9.8	−14.0	8.0	20.0	24.0	28.0		
		UV index	14,852.0	3.9	3.2	1.0	1.0	1.0	7.0	10.0		
		PM2.5 conc.	14,852.0	51.3	43.9	0.0	24.3	34.0	64.4	486.3		
		Is daytime	14,852.0								2.0	no 8037.0
		Wind direction (16 points)	14,852.0								16.0	WSW 4155.0
		Weather description	14,852.0								19.0	Partly cloudy 6065.0

3.3. Data Preprocessing and Feature Extraction

Data preprocessing is a process through which the raw data undergoes a rigorous transformation to make it understandable and implementable for the end user. Data preprocessing is essential for any analysis, as uncleaned unprocessed data will only lead to terrible results. Data quality directly influences the overall quality of information derived from results.

The collected data in this study was processed and made available for feature extraction through multiple preprocessing techniques such as missing value imputation, data cleaning, encoding of categorical features, and data scaling.

Feature engineering and selection is the part of machine learning that predominately affects the overall efficiency of any proposed model. It requires a complete understanding of the data and a thorough analysis. This analysis not only helps in understanding the raw data's features but also helps to create newer ones. In this study, the feature engineering was divided into four parts. The first part involved removing irrelevant features whose value remained the same throughout the data. These features included QC name, Loc ID, weather code, duration, and weather icon URL. The next part combined a few different features to produce a new feature, e.g., hour, day, month, and year, to form a DateTime

feature. Feature engineering also needs to resolve the problem of multicollinearity. Multicollinearity in this dataset occurred when there was a strong relationship between our dependent variable or features. It would seriously impact the overall interpretability and generalization of our model. The solution to this problem was to use Pearson correlation. The models were then fed this processed data for additional analysis and prediction.

3.4. Modelling

In recent years, neural network models such as LSTM have proven to be extremely useful in solving such time series problems. LSTM is a modified version of the recurrent neural network (RNN). Figure 5 illustrates an LSTM block. LSTM work exceptionally well in solving long-term dependency problems. The core feature of the LSTM is its cell state. The cell state is a horizontal line running on top of the multiple LSTM block, as shown in Figure 4. Cell states are similar to conveyor belts that carry information from multiple LSTM blocks. Besides the cell state, LSTM also consists of three gates, namely (1) forget gate, (2) input gate (3) output gate. The forget gate is primarily used to remove unwanted information from the cell state.

The status of the cell is updated using input gates. First, the sigmoid layer will decide which value it will update through the following equation:

$$i_t = \sigma(W_t\,[h_{t-1}, x_t] + b_i) \quad (1)$$

Then, the Tanh layer will create a vector of new candidate values to be added through the following equation:

$$C\sim_t = \tanh(W_c[h_{t-1}, x_t] + b_c) \quad (2)$$

The final cell state contains both these values, as shown in the following equation:

$$C_t = f_t \times C_{t-1} + i_t \times C\sim_t \quad (3)$$

Output gates are used for calculating the value of the next hidden state. It achieves this by passing the current hidden and previous state through the sigmoid function. The new cell state generated is also passed through the Tanh function. Then, these values are multiplied to acquire the next hidden state used for prediction. These two mathematical equations are listed as follows:

$$O_t = \sigma(W_o\,[h_{t-1}, x_t] + b_o) \quad (4)$$

$$h_t = O_t \times \tanh(C_t) \quad (5)$$

In this study, besides incorporating neural network models such as LSTM and LSTM encoder and decoder, a thorough comparative analysis was also conducted between these models and a traditional time series model such as FbProphet for the forecasting of $PM_{2.5}$.

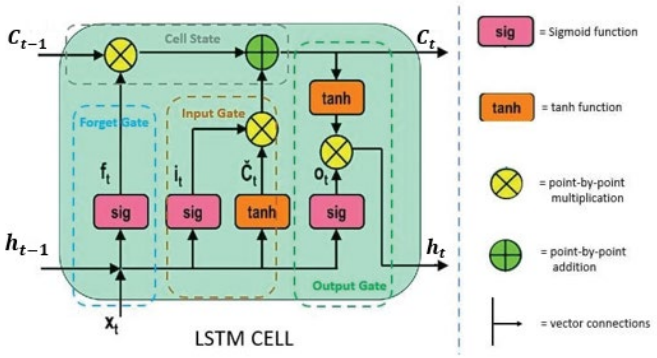

Figure 5. LSTM unit with its gates and cell states.

According to the proposed methods, this study made daily and hourly $PM_{2.5}$ concentration forecasts for Lahore, Islamabad, and Karachi for 30 days and 72 h. To achieve this goal, multivariate FbProphet, LSTM, and LSTM encoder–decoder models were utilized, and models were created separately for each city.

This study used the Keras tuner framework for optimizing hyperparameters. The parameters were selected based on the grid search mechanism. These parameters were then optimized based on the MAPE metric. This study used the parameters that achieved the lowest MAPE value for the models.

Table 2 shows the parameters used by the multivariate FbProphet model for hourly and daily forecasting. Table 3 displays the LSTM model architecture, and Table 4 illustrates the parameters used for tuning. Moreover, the time lag used for the LSTM model training was 3 and 4 for daily and hourly forecasting, respectively.

This study used the MAPE metric to compare the performance of the proposed models. MAPE is the average percentage error in any forecast. It is handy in the time series model due to its ability to stop the negative and positive errors from canceling each other out. Moreover, calculating errors in the form of percentages allows the user to conduct a better comparative evaluation. Moreover, one of this study's goals was to conduct a comparative analysis of different models; this metric appeared to be a perfect fit.

This study used LSTM encoder–decoder for sequence-to-scalar forecasting instead of conventional sequence-to-sequence forecasting. LSTM encoder–decoder was also applied to acquire the hourly and daily forecast for the $PM_{2.5}$ concentration in Lahore, Islamabad, and Karachi. The lag interval was the only difference between the daily and hourly forecast processes. The lag value was 4 for hourly forecasting and 3 for daily forecasting. Table 5 illustrates the general architecture for the final model for all three cities. Table 6 shows the parameters chosen in all the cities for hourly and daily forecasting.

Table 2. Parameters used for training of the daily and hourly multivariate FbProphet model.

	City	Forecasting	n_Changepoints	Changepoints_Prior_Scale	Seasonal_Model
1	Lahore	Daily	200	0.1	Multiplicative
		Hourly	200	0.3	Additive
2	Islamabad	Daily	200	0.1	Multiplicative
		Hourly	200	0.1	Additive
3	Karachi	Daily	100	0.3	Additive
		Hourly	100	0.2	Additive

Table 3. The model architecture of the daily and hourly LSTM model.

	Forecasting	Layer	Unit	Parameters	Dropout	Kernel Initializer	Activation Function
1	Hourly	LSTM	128	86,528		Glorot_Uniform	Tanh
		Dropout			0.2		
		Dense	100	12,900		He_Normal	Relu
		Dense	75	7575		He_Normal	Relu
		Dropout			0.2		
		Dense	50	3800		He_Normal	Relu
		Dense	25	1275		He_Normal	Relu
		Dense	1	51			
2	Daily	LSTM	128	81,408			Tanh
		Dense	75	9675			Relu
		Dense	50	3800			Relu
		Dense	1	51			

Table 4. Parameters used for training of the daily and hourly multivariate LSTM model.

	City	Forecasting	Epochs	Batch Size	Optimizer	Learning Rate	Loss
1	Lahore	Hourly	50	32	Adam	0.01	MAE
		Daily	50	8	Adam	0.01	MAE
2	Islamabad	Hourly	50	32	Adam	0.01	MAE
		Daily	50	8	Adam	0.01	MAE
3	Karachi	Hourly	50	32	Adam	0.01	MAE
		Daily	50	8	Adam	0.01	MAE

Table 5. The model architecture of the daily and hourly LSTM encoder–decoder model.

Forecasting		Layer	Unit	Activation Function
1	Daily and Hourly	LSTM	100	Tanh
		LSTM	100	Tanh
		Repeat Vector		Tanh
		LSTM	100	
		LSTM	100	Tanh
		Time Distributed	1	

Table 6. Parameters used for training of the daily and hourly multivariate LSTM encoder–decoder model.

	City	Forecasting	Epochs	Batch Size	Optimizer	Learning Rate	Loss
1	Lahore	Hourly	50	32	Adam	0.01	MAE
		Daily	50	8	Adam	0.01	MAE
2	Islamabad	Hourly	50	32	Adam	0.01	MAE
		Daily	50	8	Adam	0.01	MAE
3	Karachi	Hourly	50	32	Adam	0.01	MAE
		Daily	50	8	Adam	0.01	MAE

4. Results

This section will discuss the impact of the weather on air quality and the overall daily and hourly forecasting results for $PM_{2.5}$.

4.1. Lahore

As shown in Figure 6, Lahore had the highest $PM_{2.5}$ concentration and the most unhealthy and hazardous days compared to Karachi and Islamabad. Unlike Islamabad and Karachi, which suffered from a negligible number of hazardous days, Lahore had almost 12% of days in a year with a $PM_{2.5}$ concentration of more than 300. The primary reason for such a high concentration was due to many different weather features such as temperature, humidity, dew point, wind speed, UV index, visibility, pressure, cloud cover, and precipitation. All these features negatively correlated with the $PM_{2.5}$ concentration. A negative correlation indicated that as the value of these features decreased, the value of $PM_{2.5}$ increased. Lahore's $PM_{2.5}$ concentration strongly negatively correlated with wind speed, temperature, and dew point, as shown in Figure 7.

As mentioned in the objectives, this study made a daily forecast for the next 30 days and an hourly forecast for the next 72 h through the multivariate FbProphet, LSTM, and LSTM encoder–decoder models. The results displayed in Table 7 show that the LSTM encoder–decoder model performed the best, while multivariate FbProphet performed the worst for both hourly and daily forecasting. Multivariate FbProphet and LSTM obtained a MAPE value of 63.9% and 15.6% for hourly and 31.5% and 29.4% for daily forecasting, respectively. For daily forecasting, the LSTM encoder–decoder model achieved a MAPE

value of 28.2%, while it obtained a value of 11.75% for hourly forecasting. Figures 8–13 compare the predicted and true values for $PM_{2.5}$ concentration for the multivariate FbProphet, LSTM, and LSTM encoder–decoder models.

Table 7. Result for hourly and daily forecasting for all the proposed models in the Lahore, Islamabad, and Karachi datasets.

City	Forecasting	Models	MAPE
Lahore	Hourly	Multivariate FbProphet	63.9%
		LSTM	15.6%
		LSTM encoder–decoder	11.7%
	Daily	Multivariate FbProphet	31.5%
		LSTM	29.4%
		LSTM encoder–decoder	28.2%
Karachi	Hourly	Multivariate FbProphet	56.2%
		LSTM	7.6%
		LSTM encoder–decoder	7.4%
	Daily	Multivariate FbProphet	52.9%
		LSTM	47.2%
		LSTM encoder–decoder	42.1%
Islamabad	Hourly	Multivariate FbProphet	17.7%
		LSTM	11.6%
		LSTM encoder–decoder	9.5%
	Daily	Multivariate FbProphet	19.2%
		LSTM	15.2%
		LSTM encoder–decoder	15.1%

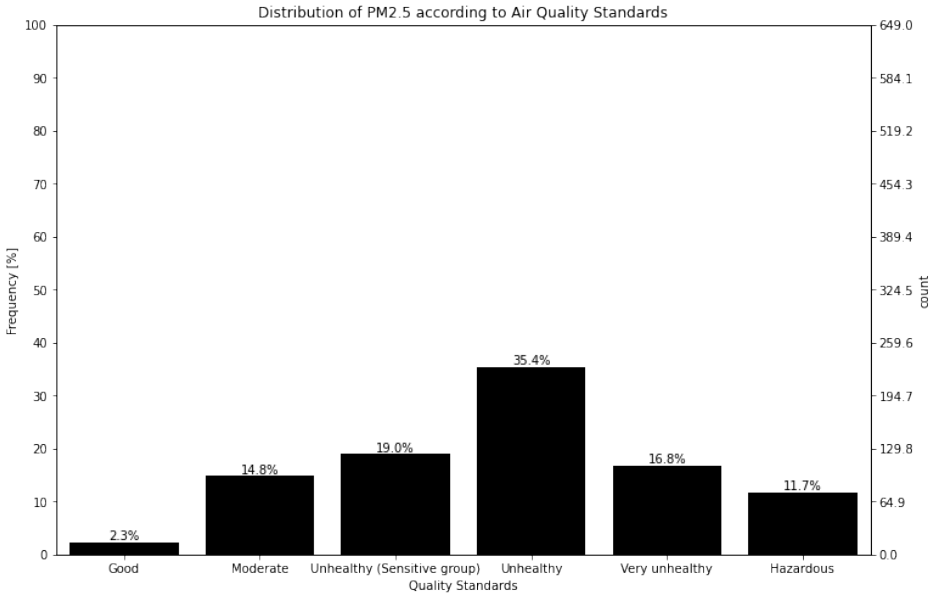

Figure 6. Depiction of the daily distribution of $PM_{2.5}$ concentrations according to EPA standards for Lahore.

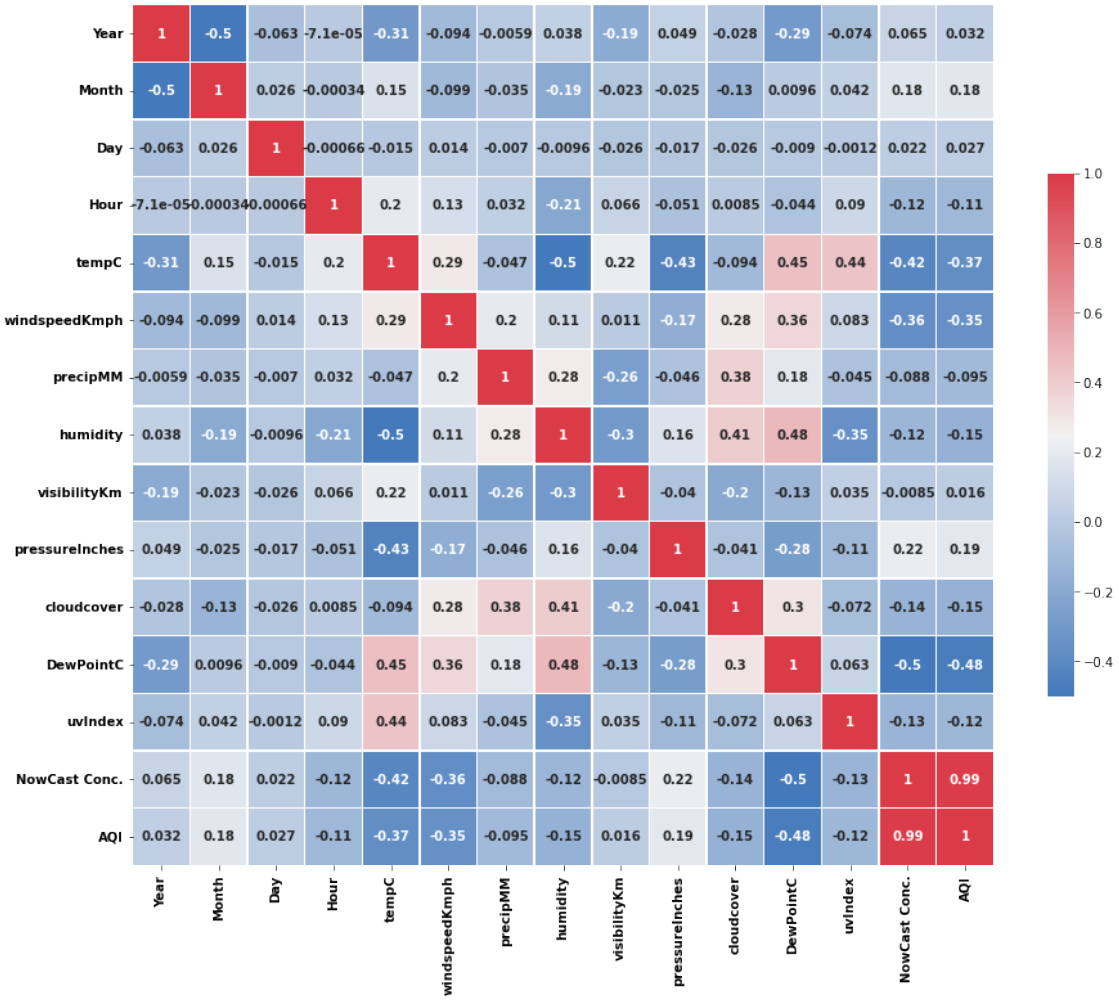

Figure 7. Correlation heat map for Lahore.

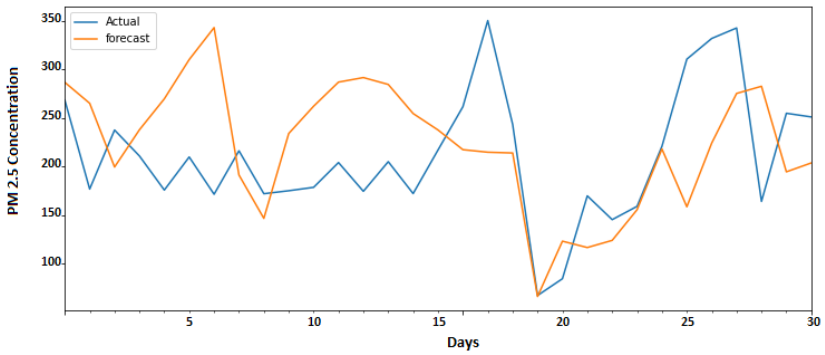

Figure 8. Forecasting for the next month for Lahore using the multivariate FbProphet model.

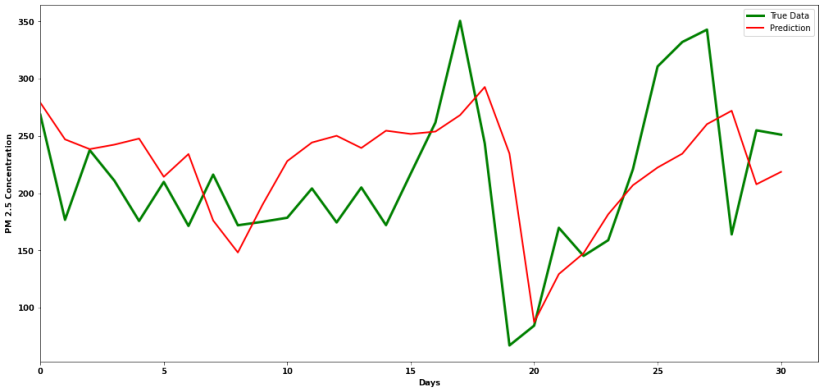

Figure 9. Forecasting for the next month for Lahore using the LSTM model.

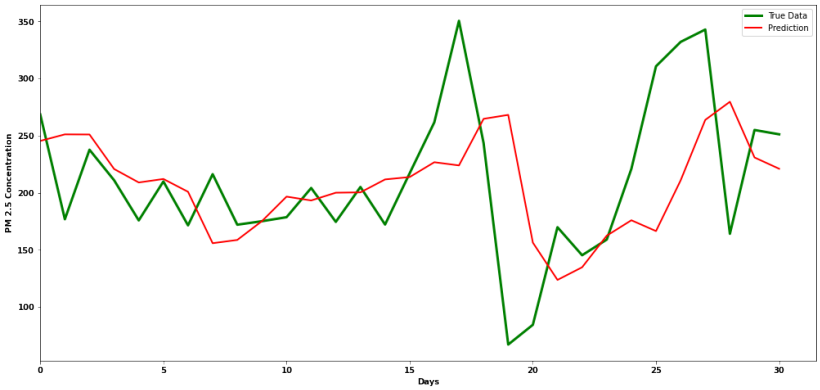

Figure 10. Forecasting for the next month for Lahore using the LSTM encoder–decoder model.

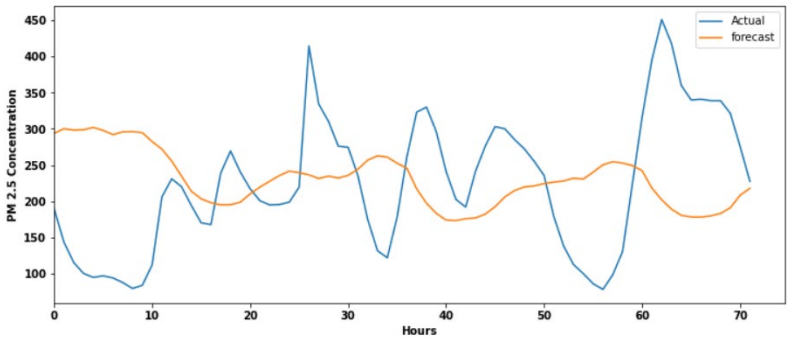

Figure 11. Forecasting for the next 72 h for Lahore using the multivariate FbProphet model.

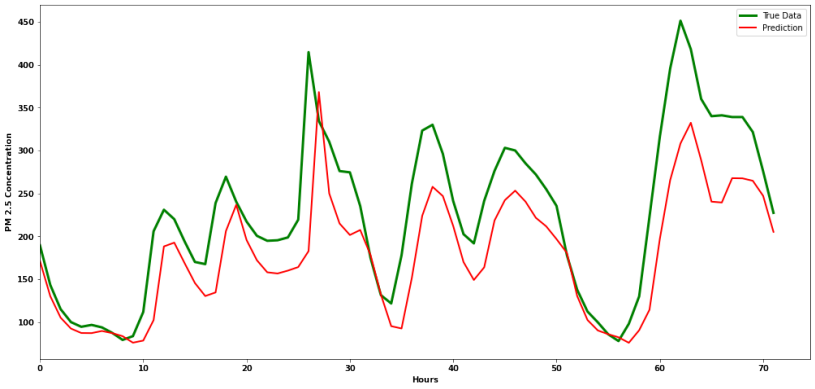

Figure 12. Forecasting for the next 72 h for Lahore using the LSTM model.

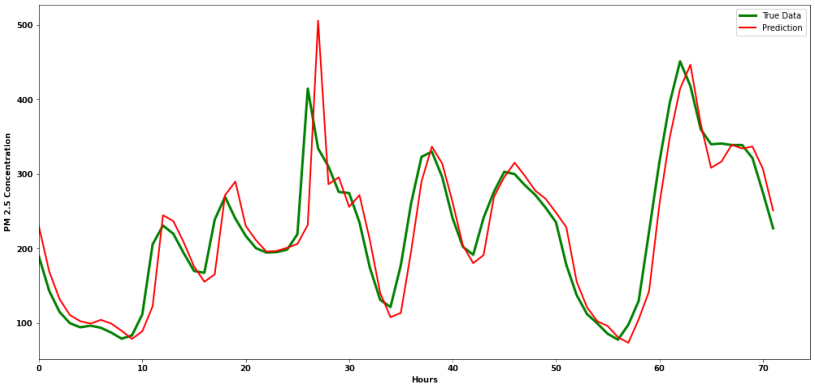

Figure 13. Forecasting for the next 72 h for Lahore using the LSTM encoder–decoder model.

4.2. Islamabad

As shown in Figure 14, Islamabad had the lowest $PM_{2.5}$ concentration and the lowest number of unhealthy and hazardous days compared to both Lahore and Karachi. These low numbers were mainly due to the overall impact of different weather features on the city. All these features had a negative correlation with the $PM_{2.5}$ concentration. Islamabad's $PM_{2.5}$ concentration strongly negatively correlated with wind speed, temperature, pressure, visibility, and dew point, as shown in Figure 15.

The results displayed in Table 7 show that the LSTM encoder–decoder model performed the best, while multivariate FbProphet performed the worst for both the hourly and daily forecasting. Multivariate FbProphet and LSTM obtained a MAPE value of 17.7% and 11.6% for hourly and 19.2% and 15.2% for daily forecasting, respectively. For daily forecasting, the LSTM encoder–decoder model achieved a MAPE value of 15.1%, while it obtained a value of 9.5% for hourly forecasting. Figures 16–21 compare the predicted and true values for $PM_{2.5}$ concentration for all of the proposed models.

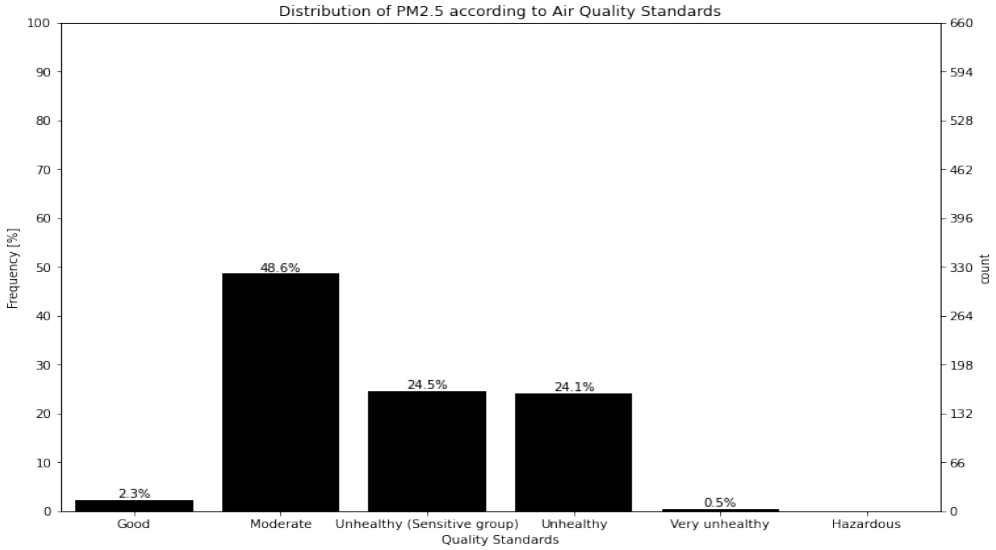

Figure 14. Depiction of the daily distribution of $PM_{2.5}$ concentrations according to EPA standards for Islamabad.

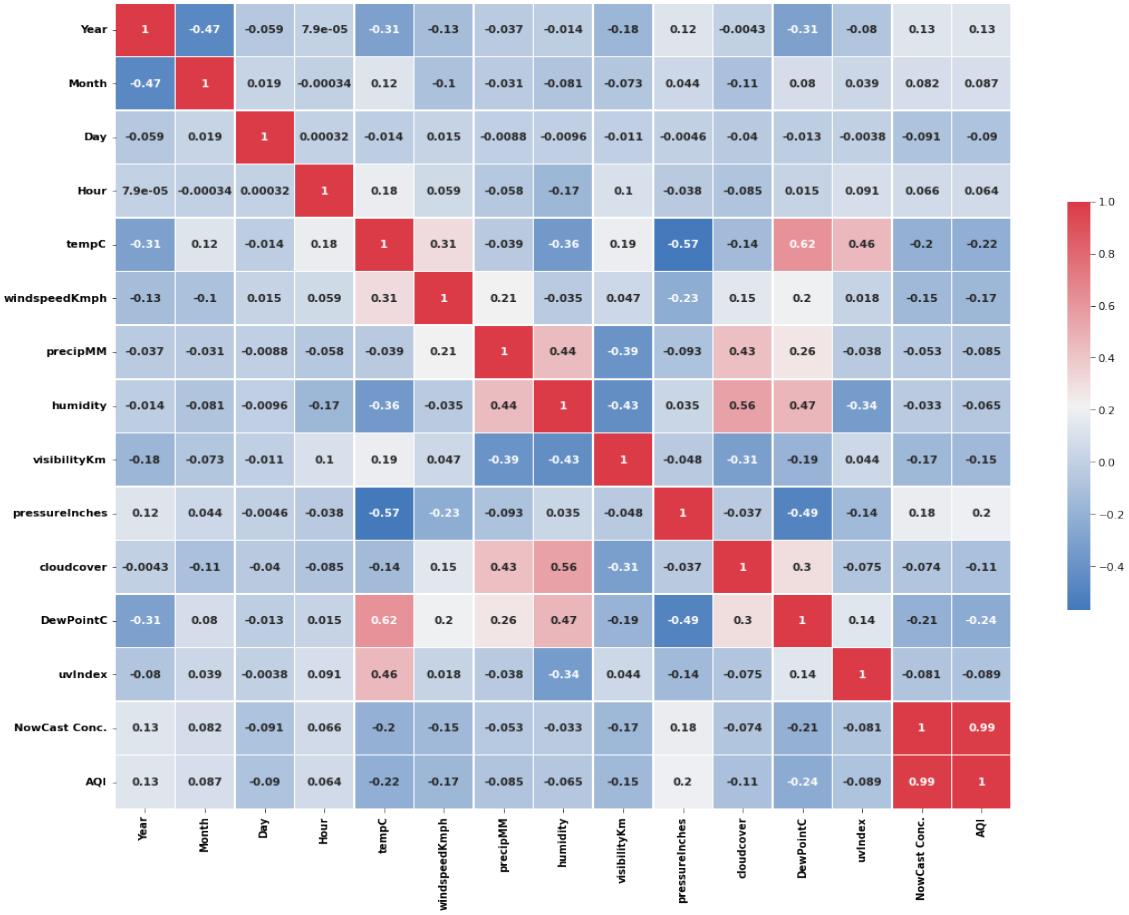

Figure 15. Correlation heat map for Islamabad.

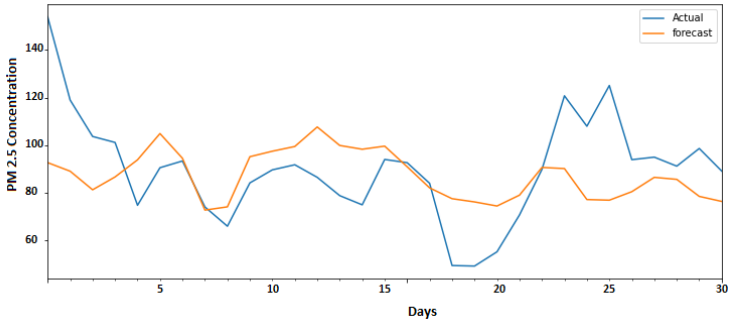

Figure 16. Forecasting for the next month for Islamabad using the multivariate FbProphet model.

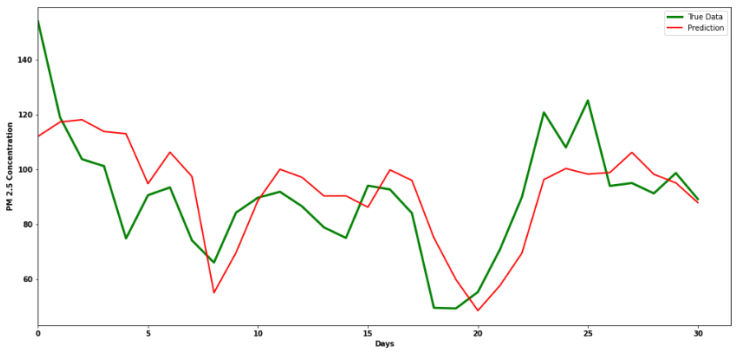

Figure 17. Forecasting for the next month for Islamabad using the LSTM model.

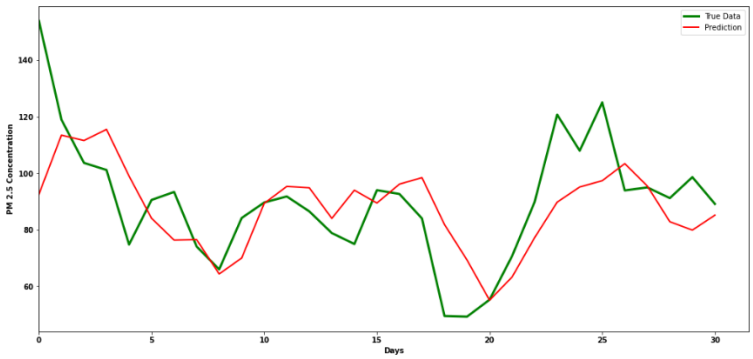

Figure 18. Forecasting for the next month for Islamabad using the LSTM encoder–decoder model.

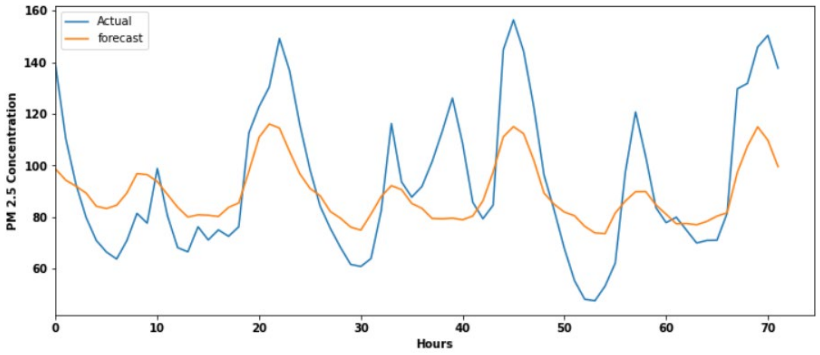

Figure 19. Forecasting for the next 72 hours for Islamabad using the multivariate FbProphet model.

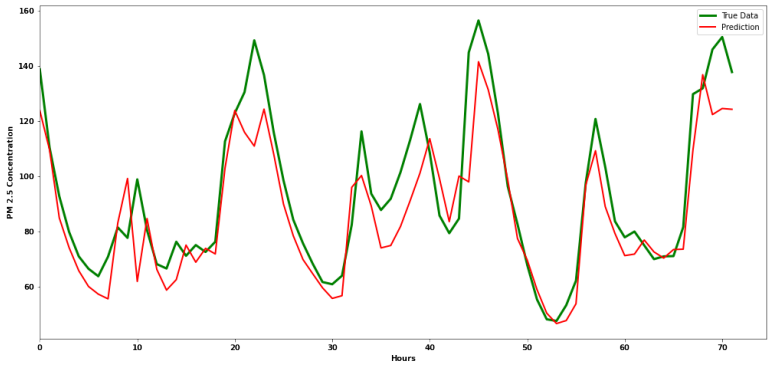

Figure 20. Forecasting for the next 72 hours for Islamabad using the LSTM model.

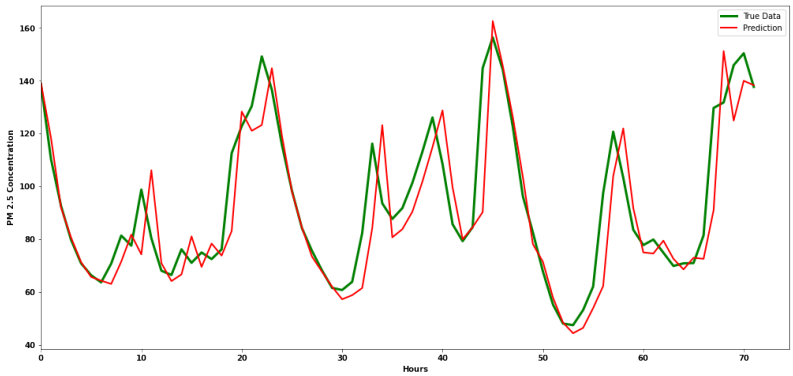

Figure 21. Forecasting for the next 72 hours for Islamabad using the LSTM encoder–decoder model.

4.3. Karachi

Figure 22 shows that Karachi had a moderate $PM_{2.5}$ concentration throughout the year. These moderate numbers were mainly due to the overall impact of different weather features on the city. All these features had a negative correlation with the $PM_{2.5}$ concentration. Karachi's $PM_{2.5}$ concentration strongly negatively correlated with wind speed, temperature, pressure, visibility, and dew point, as shown in Figure 23.

The results displayed in Table 7 show that the LSTM encoder–decoder model performed the best, while multivariate FbProphet performed the worst for both hourly and daily forecasting. Multivariate FbProphet and LSTM obtained a MAPE value of 56.2% and 7.6% for hourly and 52.9% and 47.2% for daily forecasting, respectively. For daily forecasting, the LSTM encoder–decoder model achieved a MAPE value of 42.1%, while it obtained a value of 7.4% for hourly forecasting. Figures 24–29 compare the predicted and true values for $PM_{2.5}$ concentration for all the proposed models.

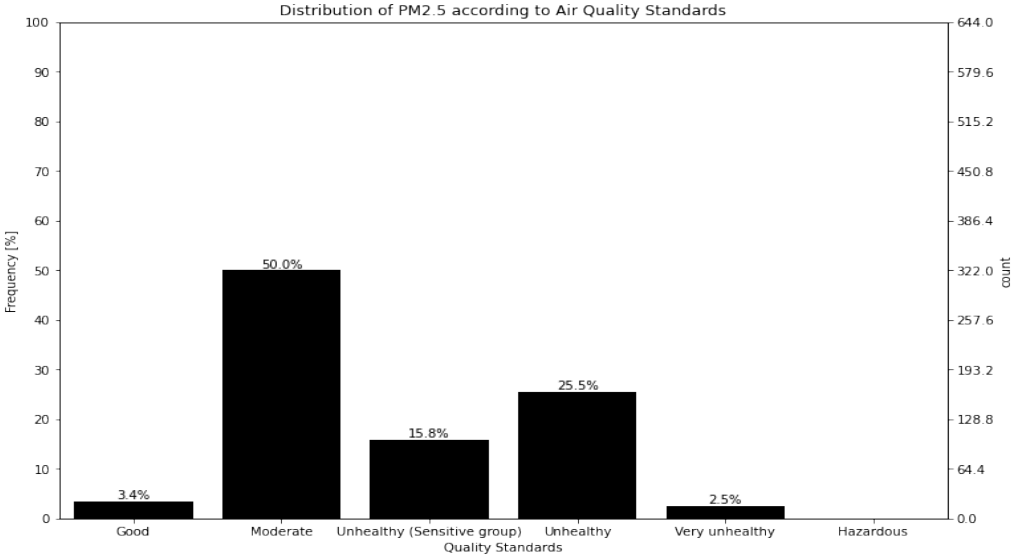

Figure 22. Depiction of the daily distribution of $PM_{2.5}$ concentration according to EPA standards for Karachi.

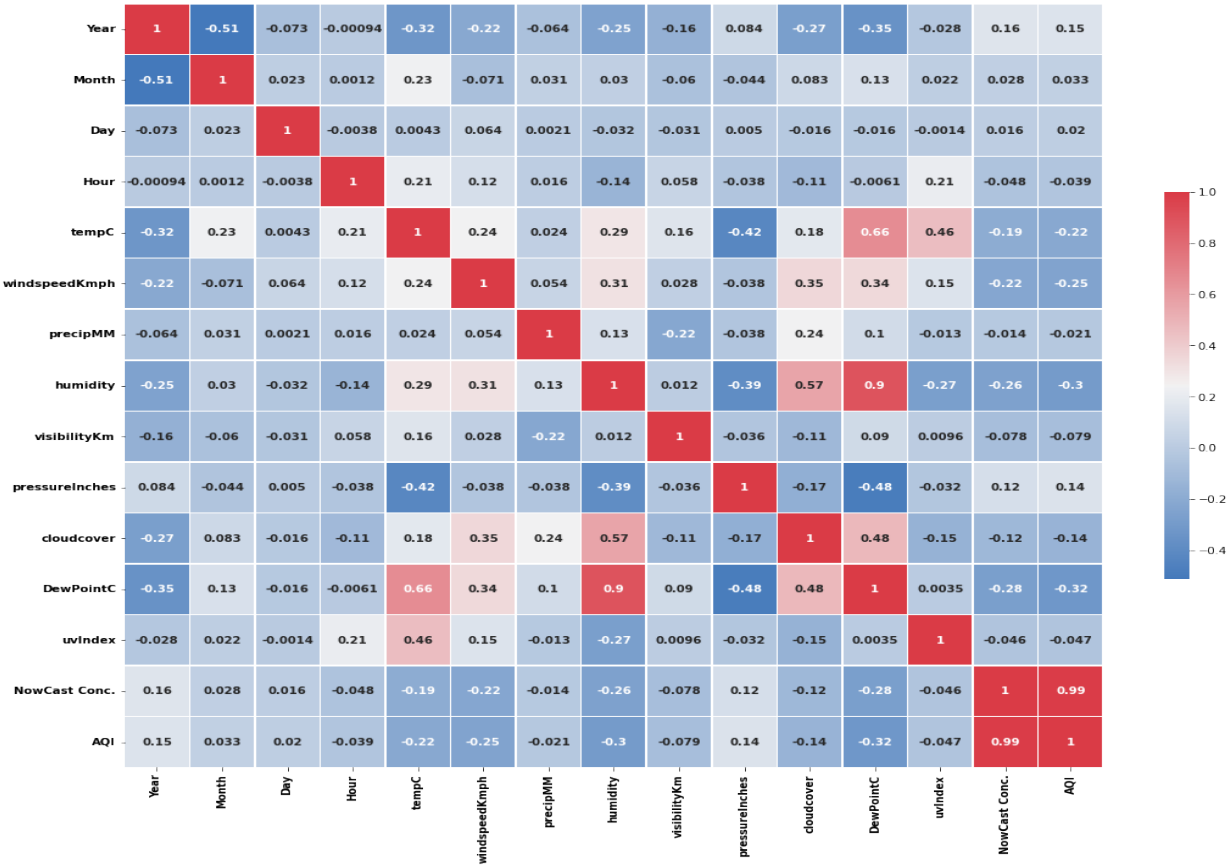

Figure 23. Correlation heat map for Karachi.

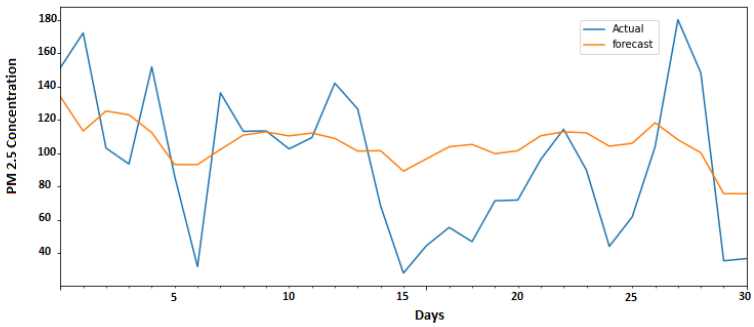

Figure 24. Forecasting for the next month for Karachi using the multivariate FbProphet model.

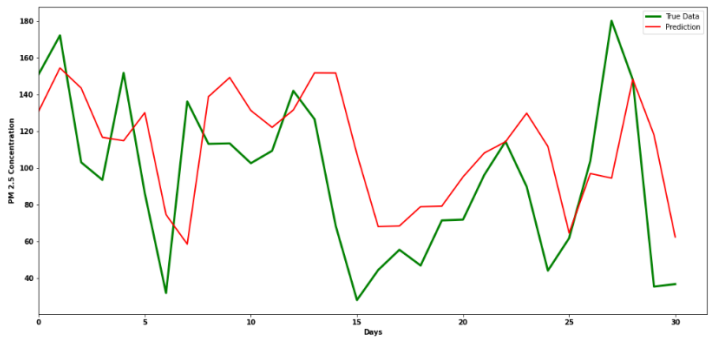

Figure 25. Forecasting for the next month for Karachi using the LSTM model.

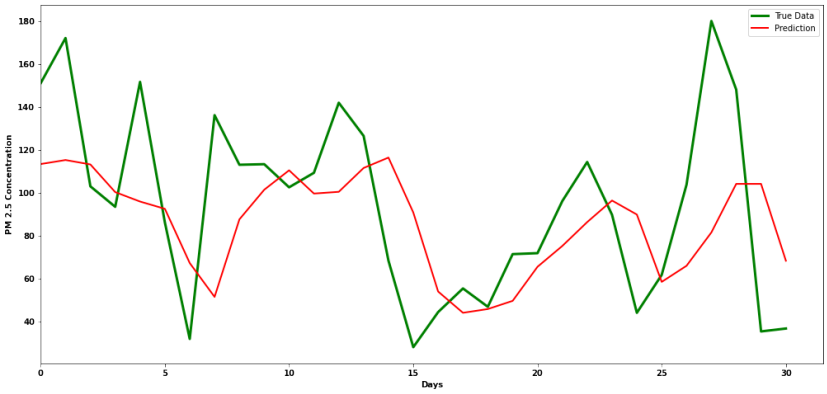

Figure 26. Forecasting for the next month for Karachi using the LSTM encoder–decoder model.

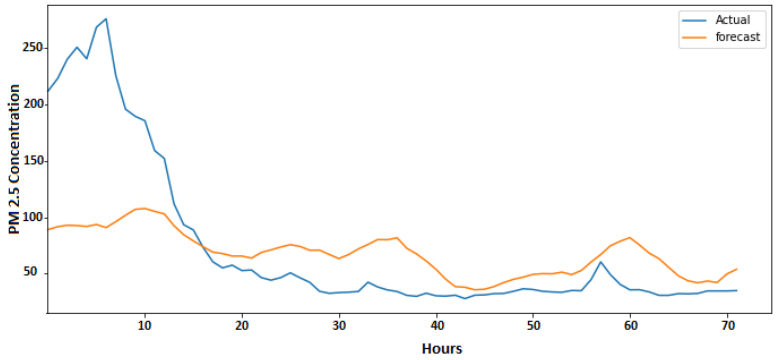

Figure 27. Forecasting for the next 72 h for Karachi using the multivariate FbProphet model.

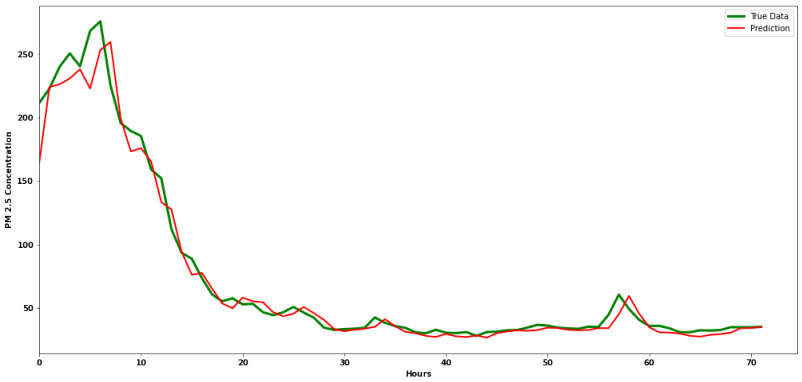

Figure 28. Forecasting for the next 72 h for Karachi using the LSTM model.

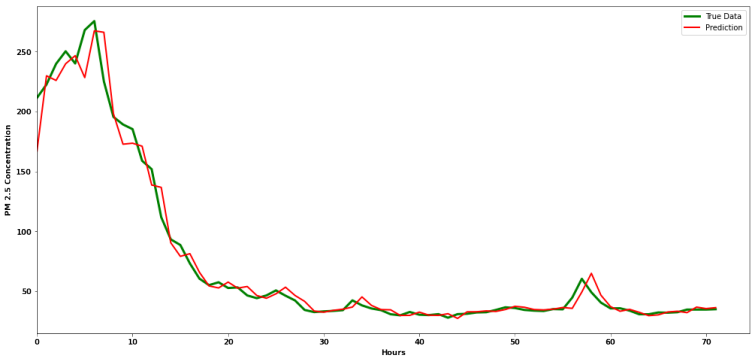

Figure 29. Forecasting for the next 72 h for Karachi using the LSTM encoder–decoder model.

5. Discussion

The weather significantly impacts the $PM_{2.5}$ concentration in different regions. A similar analysis was done by [9], where the author used penalties to find the impact of weather conditions on $PM_{2.5}$ concentration. In this analysis, they concluded that a decrease in temperature and wind speed was the primary reason behind the increase in $PM_{2.5}$ concentration. However, unlike the research conducted by [9], this research also analyzed other weather parameters such as humidity, pressure, and cloud cover. Moreover, this study concluded that besides wind speed and temperature, other parameters shown in Figures 7, 15 and 23 also had a substantial negative correlation with the $PM_{2.5}$ concentration in Pakistan.

The models proposed in this study achieved a very good MAPE value for forecasting daily and hourly $PM_{2.5}$ levels in multiple cities in Pakistan. Table 7 displays the combined results for the proposed models. These results confirmed that the prediction was better for hourly forecasting than daily forecasting. This was because hourly forecasting had more data (11,000 records) than daily forecasting (650 records). Similarly, deep learning techniques, such as LSTM and LSTM encoder–decoder, outperformed the more conventional machine learning models. This was mainly because deep learning models are more robust and flexible in handling a sudden peak in $PM_{2.5}$ concentration. Moreover, from the results, it could also be seen that the proposed multivariate models (multivariate FbProphet, LSTM, and LSTM encoder–decoder) performed better than the traditionally used univariate model. These results highlighted the importance of including more relevant features for forecasting $PM_{2.5}$ concentrations.

The author in [13] also used similar deep learning approaches. In [13], the author proposed forecasting air quality for the next 48 h by utilizing a combination of neural network models, which included ANN, CNN, and LSTM. In this study, the author conducted forecasting for Taiwanese and Chinese datasets. The author in [13] did not use weather as a feature, as their primary aim was the spatiotemporal analysis of air pollutants. In contrast, this research analyzed the data from Pakistan and included weather information as they significantly impacted the region's pollutant concentration. In [13], the authors obtained an MAE value of up to 10 for a 6-hour prediction in different regions of China. Similarly, in terms of accuracy, LSTM encoder and decoder obtained the highest accuracy among the models employed in this study. In contrast, the multivariate FbProphet model obtained the lowest accuracy. LSTM encoder and decoder obtained a MAPE value of up to 7.4% for hourly (72 h) and 15.07% for the daily forecast (30 days). Given the dataset utilized in this study and the low MAPE value, LSTM encoder–decoder was the ideal model for forecasting $PM_{2.5}$ concentrations.

This study should have used statistical tests such as the T-test, ANOVA, and F-test for a more thorough model comparison. Similarly, instead of only using MAPE, other metrics such as mean absolute error (MAE) and root mean squared error (RMSE) would have provided a more in-depth analysis of the research. These tests and metrics would have provided more conclusive evidence about whether there was any statistical difference between the results of the proposed models. However, MAPE was sufficient for explaining the general behavior for an initial model comparison. Not including statistical tests was one of the limitations of this research and would be the topic of our future research.

6. Conclusions

In conclusion, this research fulfilled all its goals and objectives. This study analyzed the impact of weather on the $PM_{2.5}$ concentration across multiple cities in Pakistan. Through an in-depth exploratory data analysis and feature engineering, this study found that all the weather parameters negatively correlated with the $PM_{2.5}$ concentration. This research also compared different machine learning models for daily and hourly forecasts

of $PM_{2.5}$ concentrations. The study proved that the LSTM encoder–decoder model performed best for this dataset. Furthermore, this study provided higher-level information to the timeseries models through a combination of pollutants and weather data. This additional information was crucial in improving the model's accuracy, as shown in Table 7, with it ranging from 15.1% to 63.9%

In the future, it is suggested to use additional data for a more in-depth analysis. Moreover, it is also suggested to add other features, such as emissions data, to acquire a more comprehensive analysis of the models. Quantitative statistical tests such as the T-test, ANOVA, and F-test should be utilized going forward in order to obtain more definitive results. Finally, the latest state-of-the-art time series models, such as the transformers and attention-based models, could also be incorporated to forecast pollutants. These models, with their ability to eliminate recurrence and parallelization, can lead to less complex and more accurate models.

Author Contributions: Conceptualization, H.M., F.A. and J.R.; methodology, K.H.W., H.M., F.A. and J.R.; software, A.M.A.-M. and M.T.; validation, H.M., A.S. and M.T.; formal analysis, K.H.W., F.A. and J.R.; investigation, K.H.W., H.M., F.A., A.S. and M.T.; resources, H.M. and A.M.A.-M.; data curation, K.H.W. and H.M.; writing—original draft preparation, K.H.W., H.M., F.A. and J.R.; writing—review and editing, F.A., A.M.A.-M., A.S., M.T. and J.R.; visualization, K.H.W., H.M. and A.M.A.-M.; supervision, F.A., A.S. and J.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The datasets used are available at [18,19].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. World Health Organization. Health Guidelines. 2020. Available online: https://www.who.int/data/gho/data/themes/theme-details/GHO/air-pollution (accessed on 1 June 2022).
2. IQAir. 2019 World Air Quality Report Region & City PM2.5 Ranking. *Air Qual. Rep.* **2019**, 1–35. Available online: https://www.iqair.com/world-most-polluted-cities/world-air-quality-report-2019-en.pdf (accessed on 1 June 2022).
3. Castelli, M.; Clemente, F.M.; Popovič, A.; Silva, S.; Vanneschi, L. A Machine Learning Approach to Predict Air Quality in California. *Complexity* **2020**, *2020*, 8049504. https://doi.org/10.1155/2020/8049504.
4. Liu, H.; Li, Q.; Yu, D.; Gu, Y. Air quality index and air pollutant concentration prediction based on machine learning algorithms. *Appl. Sci.* **2019**, *9*, 4069. https://doi.org/10.3390/app9194069.
5. Londhe, M.M. Data Mining and Machine Learning Approach for Air Quality Index Prediction. *Int. J. Eng. Appl. Phys.* **2021**, *1*, 136–153. Available online: https://ijeap.org/ijeap/article/view/28 (accessed on 1 June 2022).
6. Ghasemi, A.; Amanollahi, J. Integration of ANFIS model and forward selection method for air quality forecasting. *Air Qual. Atmos. Health* **2019**, *12*, 59–72. https://doi.org/10.1007/s11869-018-0630-0.
7. Alireza, R.; Jamil, A.; Tzanis, C.G. Air quality data series estimation based on machine learning approaches for urban environments. *Air Qual. Atmos. Health* **2021**, *14*, 191–201. https://doi.org/10.1007/s11869-020-00925-4.
8. Liaw, J.-J.; Chen, K.-Y. Using high-frequency information and rh to estimate aqi based on svr. *Sensors* **2021**, *21*, 3630. https://doi.org/10.3390/s21113630.
9. Lee, M.; Lin, L.; Chen, C.; Tsao, Y.; Yao, T.; Fei, M. Forecasting Air Quality in Taiwan by Using Machine Learning. *Sci. Rep.* **2020**, *10*, 4153. https://doi.org/10.1038/s41598-020-61151-7.
10. Ye, Z. Air Pollutants Prediction in Shenzhen Based on ARIMA and Prophet Method. *E3S Web Conf.* **2019**, *136*, 05001. https://doi.org/10.1051/e3sconf/201913605001.
11. Eknath, G.; Aniket, K.; Muley, A.; Kailas, N.; Parag, D.; Bhalchandra, U. Autoregressive integrated moving average time series model for forecasting air pollution in Nanded city, Maharashtra, India. *Model. Earth Syst. Environ.* **2018**, *4*, 1435–1444. https://doi.org/10.1007/s40808-018-0493-2.
12. Ma, J.; Ding, Y.; Cheng, J.C.P.; Jiang, F.; Tan, Y.; Gan, V.J.L.; Wan, Z. Identification of high impact factors of air quality on a national scale using big data and machine learning techniques. *J. Clean. Prod.* **2020**, *244*, 118955. https://doi.org/10.1016/j.jclepro.2019.118955.
13. Soh, P.; Chang, J.; Huang, J. Adaptive Deep Learning-Based Air Quality Prediction Model Using the Most Relevant Spatial-Temporal Relations. *IEEE Access* **2018**, *6*, 38186–38199. https://doi.org/10.1109/ACCESS.2018.2849820.
14. Zhao, Z.; Qin, J.; He, Z.; Li, H.; Yang, Y.; Zhang, R. Combining forward with recurrent neural networks for hourly air quality prediction in Northwest of China. *Environ. Sci. Pollut. Res.* **2020**, *27*, 28931–28948. https://doi.org/10.1007/s11356-020-08948-1

15. Du, S.; Li, T.; Member, S.; Yang, Y.; Horng, S. Deep Air Quality Forecasting Using Hybrid Deep Learning Framework. *IEEE Trans. Knowl. Data Eng.* **2019**, *33*, 2412–2424. https://doi.org/10.1109/TKDE.2019.2954510
16. Sethi, J.K.; Mittal, M. A new feature selection method based on machine learning technique for air quality dataset. *J. Stat. Manag. Syst.* **2019**, *22*, 697–705. https://doi.org/10.1080/09720510.2019.1609726.
17. Freeman, B.S.; Taylor, G.; Gharabaghi, B. Forecasting air quality time series using deep learning Forecasting air quality time series using deep learning. *J. Air Waste Manag. Assoc.* **2018**, *68*, 866–886. https://doi.org/10.1080/10962247.2018.1459956.
18. US Consulate Pakistan. Air Quality Data. 2021. Available online: https://www.airnow.gov/international/us-embassies-and-consulates/ (accessed on 1 June 2022).
19. World Weather Online. Weather Data. 2021. Available online: https://www.worldweatheronline.com/ (accessed on 1 June 2022).